6G－6

話者の唇の動き検出による音声対話制御の一手法**

黄 英傑*　土肥 浩*　石塚 満*

東京大学***

1　はじめに

人間とコンピュータのマンマシンインタフェースの一つの手段として、音声方式[1]が挙げられる。音声認識を行う場合に、音声情報とともに口唇画像の情報を使うと認識の性能の向上に役立つ[2]。特に雑音のある環境下で、話者の話し以外の音声が誤って認識されてしまうのを防ぐために音声認識装置に話しの始めを提示する必要がある。本文はマイク付きの小型CCDカメラで自動的に話者の唇の動きを追跡し、話しの始めと終りを検出して、音声認識装置と連動する手法を提案する。

2　システムの概要

従来の音声認識システムでは、誤って非話者の音声を認識してしまうのを防ぐために、何らかの方法、例えばフットスイッチで話しの始めを音声認識システムに提示する。擬人化エージェントシステムのように、より自然な対話環境のヒューマンインタフェースを構築するために、このような接触型のスイッチの代わりに話者の唇の動きを検出して、話者の話しの初めを音声認識システムに提示する。音声と画像を同時に入力するために、マイク付きのカメラを使用する。話者が話しをする時、自然的にマイクに近付くので、入力画像は顔の下の部分（鼻、唇、顎を含める）である。認識の結果に基づいて、ワークステーションのプリンタポートから音声認識システムに話しの始めと終りを提示する。

3　口の開閉の判断

入力画像から特定の部分を取り出す場合に、対象部分の色、形状、輝度などの情報を利用することが考えられる。ここではシステムのリアルタイム性の要求を満たすために、輝度情報を利用する。入力画像は、１６０×１２０画素、２５６階調の白黒画像である。唇の輝度値は唇の周りの肌色よりやや低いが、周辺環境の照明の影響で安定的に直接唇を抽出することは困難である。口の内部の輝度値は入力画像の中で最も低いという事実を利用して[3]、口の内部の部分を抽出し、口の開閉の判断の根拠とする。

3．1　口の内部の抽出

入力画像は顔の下の部分であるから、唇を含めて肌の輝度値を持つ画素が一番多い。まず、入力画像の輝度値の平均値と分散を取り、以下の式で閾値を求めて入力画像を２値化する。

閾値＝（輝度値の平均値＋輝度値の分散）

図１に、(a)と(b)は原画像を、(c)と(d)は２値化した画像を示している。(c)と(d)において一番大きな領域は口の内部の部分を表す。

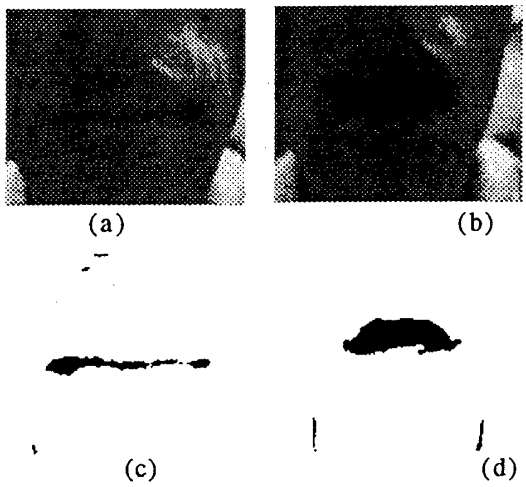

図１：　原画像(a),(b) と２値化した画像(c),(d)

---

* Ying-jieh Huang, Hiroshi Dohi, and Mitsuru Ishizuka(huang@miv.t.u-tokyo.ac.jp)

** A method of dialogue management based on lip movement detection.

*** University of Tokyo, Faculty of Engineering,7-3-1, Hongo, Bunkyo-ku, Tokyo 113, JAPAN

3．2　判別ボックス

口の開閉を判断するために、抽出した口の内部の周りに図２に示したような判別ボックスを設置する。ボックスの位置とA, B, C, D の位置は以下の手順で決める。

(1) ２値化した画像のY軸への加算投影量を計算して、投影量の最大値の位置を判定ボックスの中心位置のY座標値に設定する。

(2)(1)で求めたY座標値から出発して、X方向に口内部に属する画素を探して、もし連続する画素の数が一定の長さを超えたら口内部の長さとする。口内部の長さの１／２の位置を判別ボックスのX座標値にする。

(3)口内部の長さによって、判別ボックスのサイズと判別線A, B, C, D を決める。

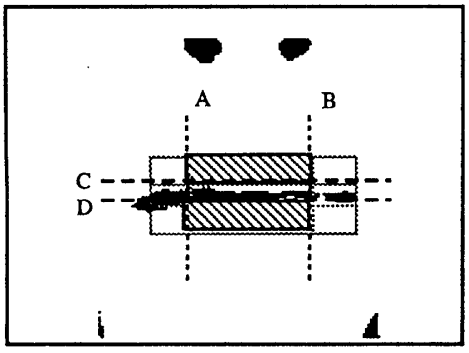

図２：判別ボックス

判定ボックスで斜線の領域に属する口内部の画素の数によって口の開閉を判断する。歯の輝度値は照明が直接当たった肌の輝度値に近いので、判定ボックスが見つからないことがあるが、この場合には口が開いていると判定する。また、演算量を削減するために、現フレームの判定ボックスの範囲を次のフレームの加算投影量計算の範囲にする。

4　実験結果

実験は普通の室内の照明の条件下で行った。実験者はマイク付きカメラから約３～１０cmぐらいの距離で話すことを仮定する。話者がカメラで撮った自分の口の映像を見ながらしゃべるので、口がマイクに向いているかを確認できる。口が映る限り、話者の頭はある程度左右に揺れたり前後に移動してもかまわない。認識率は口が閉じている時、約９７％、口が開いている時約９３％である。実行時間について、ワークステーション（　SPEC92INT=59, SPEC92FP=61）で一枚の入力画像を処理するには平均４５msかかる。

5　おわりに

本研究では、マイク付きカメラを使って話者の顔画像から口の動きを検出して、音声認識装置に話者の話し始めを提示する手法を提案した。一枚の入力画像ごとに平均45msの時間で処理できるので、約20frame/secの入力信号に対応できる。今回の実験では特別な照明を使わなかったので、話者の口が閉じるとき、唇の間の線の輝度値が上がると判別ボックスを見つけられず、口が開いたと誤認識した。一方、口が開いた時、もし上下２列の歯が同時に出ると、歯の間の黒い線を唇の間とみなして、口が閉じていると誤認識する。こういう誤りを防止するため、専用のライトを使うか、歯部分を別として処理するかは今後検討の予定である。

参考文献

[1] Yoshitaka Hiramoto, Hiroshi Dohi, Mitsuru Ishizaka: "A Speech Diaglogue Management System for Human Interface employing Visual Anthropomorphous Agent," Proc. 3rd IEEE Int' l Workshop on Robot and Human Communication, Nogoya, July 1994

[2] 田村ら：　エネルギー関数とオプティカルフローを用いた口形輪郭の抽出・補完と追跡、PRU89-20

[3] 黒田ら：　顔画像からの口部領域の自動抽出法、IE91-3